分析师：
郑兆磊
S0190520080006

宫民
S0190521040001

如何提升机构仓位测算和宏观数据预测精准度

2022 年 1 月 11 日

专题报告

报告关键点

本文介绍了卡尔曼滤波及其在投资研究中的巧妙应用。我们发现利用卡尔曼滤波可以比传统回归法更精准的测算基金行业仓位。卡尔曼滤波在宏观经济数据的建模和预测中也能够发挥重要作用。通过使用动态因子模型建模，卡尔曼滤波可以实现宏观数据共同推动因子的估计，并能够实时测算新发布数据对 GDP 等重要数据预测的信息冲击程度。

相关报告

《非线性性价比股债轮动组合》2021-08-12

《如何构建中国经济先行指数》2021-07-28

投资要点

- **卡尔曼滤波(Kalman Filter)是一种利用线性系统状态方程，通过系统输入输出的可观测数据，对系统隐含状态进行最优估计的算法，**可以从受误差影响的传感器测量中估算出最佳的实际系统状态，常用于制导与导航控制系统、计算机视觉系统和信号处理等领域。而本文创新性的发现卡尔曼滤波在投资研究中也能够产生重要作用。
- **我们巧妙的把基金的行业仓位看做隐含状态，把基金收益率当做可观测状态，然后用状态空间模型对基金收益率建模，通过卡尔曼滤波可以高效的识别出基金持仓的变动。**与传统基于线性回归的方法相比，卡尔曼滤波是一种基于贝叶斯思想的时序方法，行业平均仓位的估计误差能够减少 30%以上，重仓行业仓位估计误差减少约 50%，对于观测机构动向、指导投资具有较高价值。
- **卡尔曼滤波在宏观经济数据的建模和预测中也能够发挥重要作用。**通过使用动态因子模型建模，卡尔曼滤波可以实现宏观数据共同推动因子的估计，**并能够实时测算新发布数据对 GDP 等重要数据预测值的信息冲击程度，实现对经济数据的系统化实时监测。**
- **除了用于基金行业仓位测算和宏观数据建模，卡尔曼滤波在基金久期测算、基金经理风格识别等领域也有潜在的应用价值，**我们后续将继续探索下去。

风险提示：模型结论基于历史数据，在市场环境转变时模型存在失效的风险。

目录

1、卡尔曼滤波简介

卡尔曼滤波(Kalman Filter)是一种利用线性系统状态方程，通过系统输入输出的可观测数据，对系统隐含状态进行最优估计的算法，可以从受误差影响的传感器测量中估算出最佳的实际系统状态，常用于制导与导航控制系统、计算机视觉系统和信号处理等领域。这个方法以匈牙利数学家卡尔曼命名，而斯坦利·施密特(Stanley Schmidt)被认为是首次实际应用了卡尔曼滤波器的人。当年卡尔曼在NASA 埃姆斯研究中心访问时，施密特发现他的方法对于解决阿波罗计划的轨道预测问题很有用，后来阿波罗飞船的导航电脑就使用了这种滤波器。

除了工程领域，卡尔曼滤波在经济金融和投资研究中也能够有广泛应用。如果我们使用状态空间模型进行时间序列建模，那么卡尔曼滤波可以方便的估计隐含状态、做出预测并自动处理缺失值。举例来说，我们发现利用卡尔曼滤波可以有效及时的识别基金仓位，并且在宏观经济建模和预测中也有重要作用。

1.1、卡尔曼滤波

在具体介绍其应用场景前，我们先介绍标准卡尔曼滤波的具体推导过程。假设我们有可观测变量向量 z 和不可观测的隐含变量向量 x，两者服从一个状态空间模型（State Space Model）。在 k 时刻两者满足如下的**测量方程**（Measurement Equation）:

$$z_k = Hx_k + \eta_k \tag{1}$$

其中 η_k 为服从正态分布、协方差矩阵为 R 的白噪声过程。

隐含变量 x 自身则服从一个向量自回归过程，即状态**转移方程**如下:

$$x_k = Ax_{k-1} + e_k \tag{2}$$

其中 e_k 为服从正态分布、协方差矩阵为 Q 的白噪声过程。

在 k 时刻我们能够观测到 z_k，并且基于 k-1 期及之前的信息对隐含变量 x_k 有一个预期值 $x_{k|k-1} = E[x_k \mid F_{k-1}]$，其中 F_{k-1} 表示截止 k-1 期的可得信息。卡尔曼滤波要解决的问题就是如何权衡当前观察值和历史预期值并给出一个隐含变量的最优估计 $x_{k|k}$。一般来说，卡尔曼滤波的推导有两种思路，一种是在每一期最小化状态变量协方差矩阵的迹（Trace），另一种是从贝叶斯理论角度在每一期利用最新的观测变量信息更新得到状态变量的后验分布。这里我们给出第二种思路的推导过程。

1、 状态先验分布

因为前面我们假设状态变量 x 服从正态分布，则我们仅需要通过均值与协方差就可以描述这个分布。基于 k-1 期的信息，我们可以得到状态变量 x 的先验均值和协方差如下:

$$x_{k|k-1} = E[x_k \mid F_{k-1}] = Ax_{k-1|k-1} \tag{3}$$

$$\begin{aligned} \Sigma_{k|k-1} &= Cov[x_k, x_k \mid F_{k-1}] \\ &= Cov[Ax_{k-1} + e_k, Ax_{k-1} + e_k \mid F_{k-1}] \\ &= A\Sigma_{k-1|k-1}A^T + Q \end{aligned} \tag{4}$$

2、状态联合分布

下面我们给出 k 期状态变量 x 和可观测变量 z 的联合分布。我们将 x_k 和 z_k 放在一个向量中用 y_k 表示，则基于 k-1 期的信息，我们可以得到 y_k 服从一个联合正态分布:

$$y_k = \begin{bmatrix} x_k \\ z_k \end{bmatrix} = \begin{bmatrix} x_k \\ Hx_k + \eta_k \end{bmatrix}$$

其条件期望可表示为:

$$y_{k|k-1} = E[y_k \mid F_{k-1}] = \begin{bmatrix} x_{k|k-1} \\ Hx_{k|k-1} \end{bmatrix}$$

条件协方差可表示为:

$$\begin{aligned} \Sigma^y_{k|k-1} &= Cov[y_k, y_k \mid F_{k-1}] \\ &= \begin{bmatrix} \Sigma_{k|k-1} & \Sigma_{k|k-1}H^T \\ H\Sigma_{k|k-1} & H\Sigma_{k|k-1}H^T + R \end{bmatrix} \end{aligned}$$

3、状态后验分布

在这个步骤中我们根据 k 期新观测到的值 z_k 来调整隐含状态变量 x_k 的后验分布。首先我们给出一个引理:

引理：若随机变量 $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ 服从高斯分布，且期望为 $\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$，协方差为 $\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$，则 X_1 的条件期望和条件协方差满足如下等式:

$$\mu(x_1 \mid x_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_1 - \mu_2)$$
$$\Sigma(x_1 \mid x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

因此，给定 k 期观测变量 z_k，则状态变量 x_k 的后验条件期望和条件协方差可表示为:

$$\begin{aligned} x_{k|k} &= E[x_k \mid z_k, F_k] \\ &= x_{k|k\text{-}1} + \Sigma_{k|k-1}H^T(H\Sigma_{k|k-1}H^T + R)^{-1}(z_k - z_{k|k-1}) \end{aligned}$$

$$\begin{aligned} \Sigma_{k|k} &= Cov[x_k, x_k \mid z_k, F_k] \\ &= \Sigma_{k|k-1} - \Sigma_{k|k-1}H^T(H\Sigma_{k|k-1}H^T + R)^{-1}H\Sigma_{k|k-1} \end{aligned}$$

令:

$$K_k = \Sigma_{k|k-1}H^T(H\Sigma_{k|k-1}H^T + R)^{-1} \tag{5}$$

则上述等式可化简为:

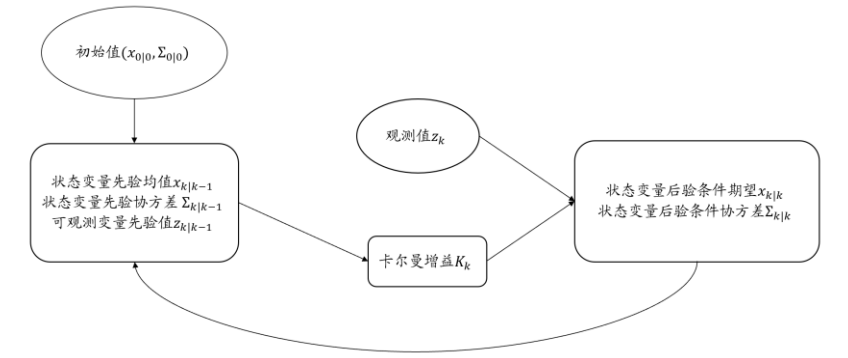
$$x_{k|k} = x_{k|k\text{-}1} + K_k(z_k - z_{k|k-1}) \tag{6}$$

$$\Sigma_{k|k} = \Sigma_{k|k-1} - K_k H\Sigma_{k|k-1} \tag{7}$$

这里 K_k 就是著名的卡尔曼增益（Kalman Gain）项，可以看到状态变量的后验期望等于先验期望 $x_{k|k\text{-}1}$ 与新数据超预期部分 $z_k - z_{k|k-1}$ 的加权和。通过简单的变换我们还可以把等式(4)表示为 $x_{k|k} = (A^{-1} - K_k)z_{k|k\text{-}1} + K_k z_k$，从这里我们可以更清晰的看出后验期望 $x_{k|k}$ 也是新观测数据与上期预期数据的加权和。

公式（3）至（7）就是卡尔曼滤波计算所需的 5 条全部等式。其中（3）（4）为预测等式（Prediction），表示如何根据上期信息得到当期状态变量的先验分布。（5）（6）（7）为更新等式（Update），表示如何根据当期新观测值更新状态变量的后验分布。在实际操作中，我们首先给定 0 期的状态变量初始值 x_0 和 Σ_0，然后循环使用公式（3）-（7）进行预测和更新操作。

图表 1、卡尔曼滤波计算流程示意图



资料来源：兴业证券经济与金融研究院绘制

1.2、状态约束卡尔曼滤波

前面我们介绍了标准卡尔曼滤波的推导过程和算法，它能够有效及时的识别出隐含状态变量。对于线性动态系统，卡尔曼滤波给出的就是方差最小的状态估计值。不过，当我们对系统的状态变量有更多外生信息时，比如满足某些等式或不等式约束，那么通过利用这部分信息就有可能得到更优的结果。

1. 等式约束

首先我们来看如何将等式约束纳入卡尔曼滤波的估计过程。假设我们依然有线性动态系统满足等式（1）（2）。但是在任意 k 时刻，状态变量 x 满足如下等式约束:

$$Dx = d$$

在这种情形下，我们如何估计 k 期的状态变量 x 呢？首先我们在无约束条件下计算得到状态变量预测值 $x_{k|k}$ 和协方差矩阵 $\Sigma_{k|k}$。然后将无约束变量投影到约束空间中，即求解如下最优化问题:

$$\begin{cases} \tilde{x}_{k|k} = \arg\min_x (x - x_{k|k})^T W (x - x_{k|k}) \\ s.t. \qquad\qquad Dx = d \end{cases}$$

其中 W 是正定的加权矩阵。上述优化问题有解析解如下:

$$\tilde{x}_{k|k} = x_{k|k} - W^{-1}D^T(DW^{-1}D^T)^{-1}(Dx_{k|k} - d)$$

这里若我们设定 $W = (\Sigma_{k|k})^{-1}$，则得到的是最大似然估计量；若设定 $W = I$，则得到的是最小二乘估计量。令 $\Gamma = W^{-1}D^T(DW^{-1}D^T)^{-1}$，则上式可表示为:

$$\tilde{x}_{k|k} = x_{k|k} - \Gamma(Dx_{k|k} - d)$$

当状态变量满足等式约束时，意味着状态变量协方差矩阵不再是满秩的，其形式也需要发生相应的改变。首先我们有：

$$\begin{aligned} x_k - \tilde{x}_{k|k} &= x_k - x_{k|k} + \Gamma(Dx_{k|k} - d - (Dx_k - \mathrm{d})) \\ &= x_k - x_{k|k} + \Gamma(Dx_{k|k} - Dx_k) \\ &= (I - \Gamma D)(x_k - x_{k|k}) \end{aligned}$$

根据协方差矩阵定义我们有：

$$\begin{aligned} \tilde{\Sigma}_{k|k} &= E[(x_k - \tilde{x}_{k|k})(x_k - \tilde{x}_{k|k})^T] \\ &= E[(I - \Gamma D)(x_k - x_{k|k})(x_k - x_{k|k})^T (I - \Gamma D)^T] \\ &= (I - \Gamma D)\Sigma_{k|k}(I - \Gamma D)^T \end{aligned}$$

2. 不等式约束

假设在任意 k 时刻，状态变量 x 满足如下不等式约束：

$$D^{ieq}x \le d^{ieq}$$

在这种情形下如何估计 k 期的状态变量 x 呢？我们的思路依然是通过投影方法将预测值映射到约束空间中。首先我们在无约束条件下计算得到状态变量预测值 $x_{k|k}$。然后将无约束变量投影到约束空间中，即求解如下最优化问题：

$$\begin{cases} \tilde{x}_{k|k} = \arg\min_x (x - x_{k|k})^T W (x - x_{k|k}) \\ s.t. \qquad\qquad D^{ieq}x \le d^{ieq} \end{cases}$$

对于一般的不等式约束问题我们无法得到解析解，但可以使用数值方法求解得到 $\tilde{x}_{k|k}$。下面我们需要计算协方差矩阵 $\tilde{\Sigma}_{k|k}$，但是不等式约束条件下我们无法得到其解析形式。幸运的是，我们可以将不等式约束问题转化为等式约束问题，从而方便的解决问题。

具体来说，我们首先使用数值方法求解不等式约束问题得到 $\tilde{x}_{k|k}$，然后判断不等式约束条件中有哪些是紧的，即找到所有满足 $D_i^{ieq}x = d_i^{ieq}$ 的 i，并将其构建为一个等式约束 $D^{eq}x = d^{eq}$。也就是说我们可以把上述不等式约束问题等价转化为下述等式约束问题。这样一来，我们就可以直接利用 1.1 部分的结论得到协方差矩阵 $\tilde{\Sigma}_{k|k}$ 了。

$$\begin{cases} \tilde{x}_{k|k} = \arg\min_x (x - x_{k|k})^T W (x - x_{k|k}) \\ s.t. \qquad\qquad D^{eq} x = d^{eq} \end{cases}$$

2. 卡尔曼滤波与基金仓位测算

公募基金作为资本市场的重要参与者，它们的投资交易行为受到市场广泛关注。其持仓变动能够反映机构投资者对于市场的观点变化，对于指导投资构建交易策略具有很高的价值。但是基金的持仓信息每年只披露 4 次，完整信息更是只有半年报和年报中才有披露，因此我们需要使用别的方法获取基金仓位的高频估计值。我们发现基金仓位的测算问题可以完美转化为隐含状态变量的识别问题，从而可以使用卡尔曼滤波方法直接求解，并且估计结果更加精准计算速度也更快。

2.1、传统行业仓位测算方法

目前业界已有的基金仓位测算方法大都是基于回归的，并通过 Lasso 或 PCA 等方法来减少多重共线性的问题，这些方法都是使用过去一段时间的数据进行测算，其逻辑上得到的是这段时间的平均仓位，而不是当前时点的仓位。在下面具体介绍卡尔曼滤波方法前，我们先对我们团队过去使用的传统方法做简要回顾。

2.1.1、Lasso 回归

基金行业仓位测算的基本方法是将基金净值增长率对行业指数收益率和国债指数收益率进行带有约束条件的多元线性回归。我们可以采取 Lasso 回归法来避免多重共线性的问题，回归方程如下：

$$y = \alpha + a_0 u + a_1 x_1 + a_2 x_2 + \cdots + a_{28} x_{28}$$

然后我们求解如下最优化问题：

$$min_w \sum_{i=1}^{m} (y_i - a_i x_i)^2 + \lambda \|a\|_1$$

$s.t.\quad 0 \le a_i \le 1,\ i = 0,1,2,\ldots,28$

$a_0 + a_1 + a_2 + \cdots + a_{28} \le 1$

$60\% \le a_1 + a_2 + \cdots + a_{28} \le 95\%$，当基金为偏股混合型

$60\% \le a_1 + a_2 + \cdots + a_{28} \le 95\%$，当基金为普通股票型且时间点在 2015/8/8 前

$80\% \le a_1 + a_2 + \cdots + a_{28} \le 95\%$，当基金为普通股票型且时间点在 2015/8/8 后

其中 y 为基金复权单位净值的日度增长率，$x_1, x_2, \ldots, x_{28}$分别为 28 个申万一级行业指数的日度收益率，u 为中债国债总财富指数（CBA00601.CS）的日度收益率。

α代表现金部分的收益，$a_1, a_2, \ldots, a_{28}$分别代表各个行业的股票市值占基金净值的比例，对$a_1, a_2, \ldots, a_{28}$进行归一化处理，得到基金的行业仓位$w_1, w_2, \ldots, w_{28}$：

$$w_i = a_i / (\sum_{j=1}^{28} a_j) \ , \ i = 0,1,2,\ldots,28$$

通过添加惩罚项，Lasso 回归可以使某些行业指数收益率前的系数变为 0，从而基金收益率可以表示为一些较为重要的行业指数收益率的线性组合，这也较为符合基金持仓的实际状况，部分基金确实在一些行业上的仓位为 0。我们将惩罚系数λ设为3×10^{-6}。

2.1.2、PCA 回归

行业指数的收益率之间具有一定的相关性，回归自变量存在多重共线性，对回归结果可能产生不利影响，我们也可以采用主成分分析（PCA），从多个行业指数收益率中提取出若干相互正交的主成分，用基金净值收益率对主成分进行多元回归，然后将主成分前的系数还原为原指数收益率变量前的系数，得到行业仓位。具体过程如下。

对 28 个申万一级行业指数收益率x_i进行主成分分析，在总解释力度为 95%的情况下选择主成分，通常选到 9 个主成分，设 9 个主成分为$z_1, z_2, \ldots, z_9$，设主成分z_i与原变量x_i之间的关系为：

$$\begin{cases} z_1 = c_{1,1}x_1 + c_{1,2}x_2 + \cdots + c_{1,28}x_{28} + \alpha_0 \\ z_2 = c_{2,1}x_1 + c_{2,2}x_2 + \cdots + c_{2,28}x_{28} + \alpha_0 \\ \cdots \\ z_9 = c_{9,1}x_1 + c_{9,2}x_2 + \cdots + c_{9,28}x_{28} + \alpha_0 \end{cases}$$

将基金净值增长率 y 对 9 个主成分z_i、国债指数收益率 u 进行带有约束条件的多元线性回归，以最小化残差平方和为目标，回归方程以及约束条件如下：

$$y = d_1 z_1 + d_2 z_2 + \cdots + d_9 z_9 + d_0 u + \alpha$$

$$s.t. \quad 0 \le a_i \le 1, \ i = 0,1,2,\ldots,28$$

$$a_0 + a_1 + a_2 + \cdots + a_{28} \le 1$$

$60\% \le a_1 + a_2 + \cdots + a_{28} \le 95\%$，当基金为偏股混合型

$60\% \le a_1 + a_2 + \cdots + a_{28} \le 95\%$，当基金为普通股票型且时间点在 2015/8/8 前

$80\% \le a_1 + a_2 + \cdots + a_{28} \le 95\%$，当基金为普通股票型且时间点在 2015/8/8 后

其中，

$$(a_1, a_2, \ldots, a_{28}) = (d_1, d_2, \ldots, d_9) \begin{pmatrix} c_{1,1} & c_{1,2} \cdots c_{1,28} \\ c_{2,1} & c_{2,2} \cdots c_{2,28} \\ & \cdots \\ c_{9,1} & c_{9,2} \cdots c_{9,28} \end{pmatrix} \quad (A)$$

其中 y 为基金复权单位净值的日度增长率，$x_1, x_2, \ldots, x_{28}$分别为 28 个申万一级行业指数的日度收益率，u 为中债国债总财富指数（CBA00601.CS）的日度收益率。

得到 9 个主成分之前的系数$d_1, d_2, \ldots, d_9$之后，根据公式（A）将其转换为原行业指数收益率之前的系数$a_1, a_2, \ldots, a_{28}$，其分别代表各个行业的股票市值占基金净资产的比例，对$a_1, a_2, \ldots, a_{28}$进行归一化处理，得到基金的行业仓位$w_1, w_2, \ldots, w_{28}$：

$$w_i = a_i / (\sum_{j=1}^{28} a_j) \ , \ i = 0,1,2,\ldots,28$$

2.1.2、Lasso+PCA 回归

在我们团队的基金周报中使用的测算方法是前两种方法的改进版本。我们对主成分回归与 Lasso 回归的单只基金测算误差进行分析发现: 对于主成分回归法，主题型基金的测算误差明显大于非主题型基金；对于 Lasso 回归法，主题型基金的测算误差小于非主题型基金的测算误差。因此，对于主题型基金，我们使用 Lasso 回归法；对于非主题型基金，我们使用主成分回归法。

主题型基金与非主题型基金的判断标准设置如下:

（1）根据基金定期报告（半年报或年报）公布的持仓情况，**若该基金前 5 大持仓行业占比之和大于 70%，且持仓行业数小于 15，则认为该基金为主题型基金**;

（2）不满足以上条件的其余基金为非主题型基金。

2.2、卡尔曼滤波估算基金行业仓位

下面我们介绍如何利用卡尔曼滤波估算基金的行业仓位。我们假设基金持仓组合由若干备选指数构成。然后我们用状态空间模型对基金收益率 R_k 和备选指数收益率 r_k 建模，并设定基金在各个备选指数上的权重 w_k 为隐含状态变量，则有以下等式:

$$\begin{cases} R_k = r_k w_k + \eta_k, \eta_k \sim N(0, R) & (8) \\ w_k = w_{k-1} + e_k, e_k \sim N(0, Q) & (9) \end{cases}$$

其中等式（8）为测量方程，它表示基金净值收益率等于备选指数收益率与权重向量的加权和。等式（9）为转移方程，它表示基金的组合权重向量服从一个随机游走过程。这个模型与等式（1）（2）代表的模型最大不同之处在于测量方程中的测量矩阵 H 在这里是时变的收益率 r_k。

在运行卡尔曼滤波算法前，我们还需要给定模型的初始值即基金的初始权重

向量。我们可以将基金某一期半年报披露的权重分布作为状态变量初始值，这样能够充分利用可得信息，使得卡尔曼滤波能够更快的收敛到真实值。然后给定协方差矩阵 Q 和 R 后我们便可以开始运行卡尔曼滤波，得到每一期的持仓权重向量 w_k。需要注意的是，这里我们没有对 w_k 做任何约束，它可以取正值或负值。但由于我们把 w_k 当做组合权重，那么它需要满足一定约束，例如:

$$\begin{cases} w_k \geq 0 \\ \sum_{i=1}^{N} w_{i,k} = 1 \end{cases}$$

为了使得估算结果符合约束条件，我们需要对滤波得到的状态变量 w_k 进行调整。这里我们给出两种调整方法:

1. 滤波后加约束（后约束 KF）

这种方法直接使用无约束卡尔曼滤波进行计算得到状态变量序列 w_k，然后我们对每一期的状态变量进行调整:

a. 将所有小于 0 的权重填充为 0;

b. 将向量归一化，保证权重之和为 1。

2. 滤波中加约束（中约束 KF）

这里我们使用状态约束卡尔曼滤波进行计算，即求解如下模型:

$$\begin{cases} R_k = r_k w_k + \eta_k, \eta_k \sim N(0,R) \\ w_k = w_{k-1} + e_k, e_k \sim N(0,Q) \\ s.t. \quad w_k \geq 0 \\ \quad\quad \sum_{i=1}^{N} w_{i,k} = 1 \end{cases}$$

通过迭代计算，模型将直接输出满足约束条件的状态变量序列 w_k。

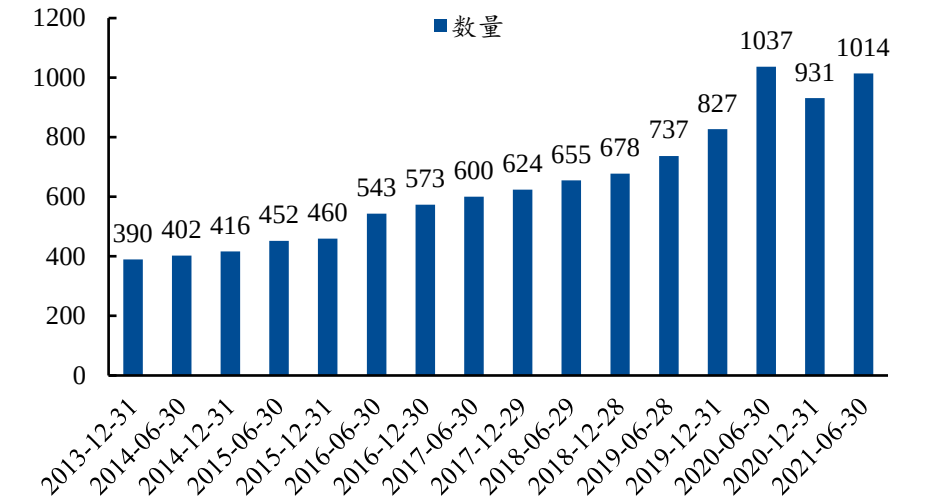
2.3、实证分析

2.3.1、数据说明

本文研究的基金样本为所有普通股票型和偏股混合型基金，只取开放式基金和初始基金，并剔除定期开放基金以及沪港深基金等不完全投资于 A 股市场的基金。在每个季度末选取成立期满两个季度、当时未到期、且规模大于 5000 万元的基金。测试时间段为 2014 年 6 月 30 日至 2021 年 6 月 30 日。下方图表给出了每

期入选的基金数量。

图表 2、每期入选的基金数量（单位：只）



资料来源：Wind，兴业证券经济与金融研究院整理

我们使用申万一级行业指数作为行业指数的代表，并且在使用卡尔曼滤波法时用中债总财富总值指数（CBA00301.CS）和货币市场基金指数（885009.WI）来代表基金组合中的非股票资产。

在测算过程中，我们分别使用上述几种方法对每只基金在二季度和四季度末的基金仓位进行计算，从而得到单只基金以及全市场基金的平均行业仓位情况，并与半年报和年报的实际披露数据进行对比，得到测算误差。

2.3.2、误差比较

我们首先定义**基金行业仓位误差**为单只基金各行业仓位测算值与实际值的离差均值，它衡量了单只基金每个行业的平均误差大小。我们再定义**行业平均仓位误差**为所有基金的测算平均行业仓位与实际平均行业仓位的离差均值，它衡量了整个基金市场平均行业仓位测算的误差大小。

下方图表给出了从 2014 年到 2021 年样本期内，各期测算误差的均值。可以看到，基于卡尔曼滤波的两种方法明显降低了对单只基金的仓位估计误差，其中后约束 KF 方法能够实现单只基金平均 2.82%的行业估计误差，而之前最好的 Lasso+PCA 方法误差有 3.21%。从基金行业平均仓位的估计误差看，基于后约束 KF 的方法能够实现 0.99%的行业估计误差，比 Lasso+PCA 方法降低了 30%以上。另外，我们发现中约束 KF 方法的表现弱于后约束 KF 方法，对于估计误差的降低效果有限。这可能是由于在卡尔曼滤波迭代过程中加入约束并没有提供有效的外生信息，反而破坏了原迭代过程的最优特征。

图表 3、基金行业仓位误差均值

测算方法	基金行业仓位误差均值
后约束 KF	2.82%
中约束 KF	3.19%
Lasso	3.45%
PCA	3.48%
Lasso+PCA	3.21%

资料来源：Wind，兴业证券经济与金融研究院

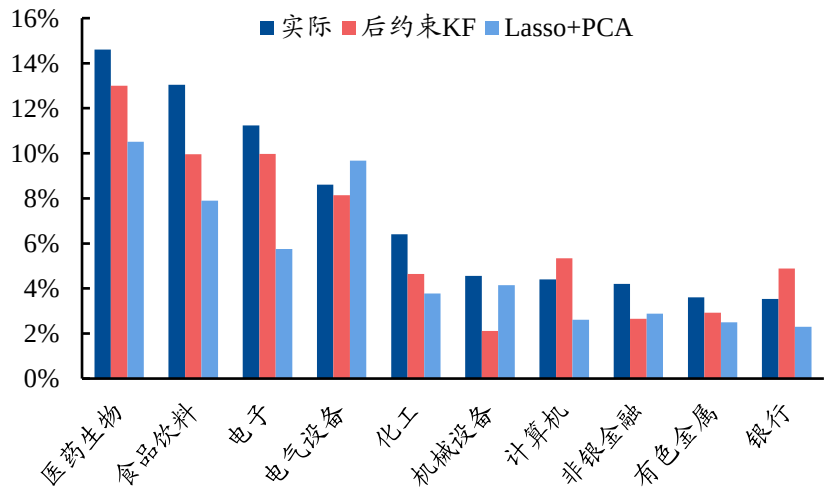
图表 4、各行业平均仓位误差均值

测算方法	行业平均仓位误差均值
后约束 KF	0.99%
中约束 KF	1.47%
Lasso	1.45%
PCA	1.78%
Lasso+PCA	1.42%

资料来源：Wind，兴业证券经济与金融研究院

下面我们以 2020 年 12 月 31 日为例，观察两种相对最优方法的测算结果与实际平均仓位的关系。从下图可以看到，2020 年底普通股票型和偏股混合型基金前三大重仓行业分别为医药生物、食品饮料和电子。直观的看，基于后约束卡尔曼滤波方法估计的平均仓位与真实仓位非常接近，显著优于传统的 Lasso+PCA 方法。

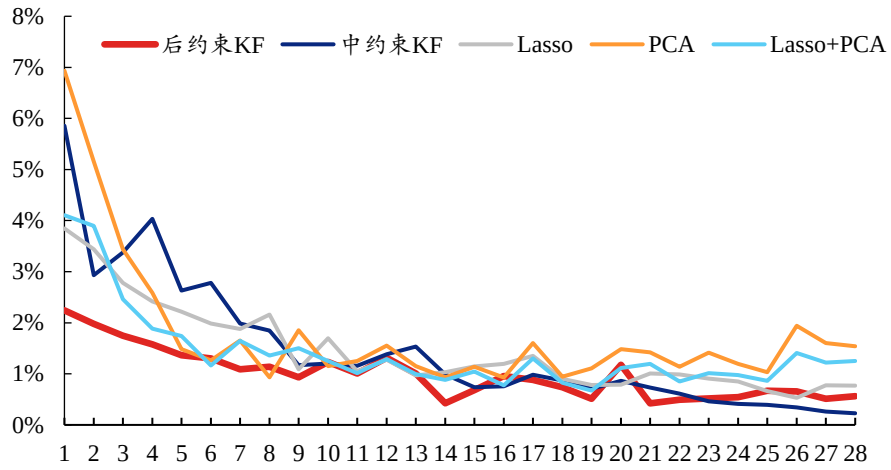
图表 5、平均仓位前十大行业测算结果（2020 年 12 月 31 日）



资料来源：Wind，兴业证券经济与金融研究院

进一步，我们还希望研究不同测算方法对于配置比例较高和较低行业的测算结果有何特征。我们分别测算每个报告期的基金平均行业仓位及误差，然后按照当期实际行业仓位降序排列，最后对所有报告期取均值。这样我们就得到不同重仓程度行业的估计误差。下方图表给出了统计结果，标红加粗的线代表后约束 KF 方法。可以看到，相比于传统回归法，后约束 KF 方法几乎在每个行业上都有更低的误差，并且在前几大重仓行业中误差有极为显著的降低。实际上，后约束 KF 方法在前三大行业的估计误差分别只有 2.24%、1.98%和 1.74%，而回归法中表现最好的 Lasso+PCA 方法误差分别达到 4.11%、3.89%和 2.46%。也就是说卡尔曼滤波能够减少 40%-50%的重仓行业估计误差，表现非常优异。

图表 6、行业平均仓位测算误差均值（按照重仓程度降序）



资料来源：Wind，兴业证券经济与金融研究院

图表 7、不同方法在前大三重仓行业的估计误差

	后约束 KF	中约束 KF	Lasso	PCA	Lasso+PCA
1	2.24%	5.86%	3.85%	6.93%	4.11%
2	1.98%	2.93%	3.44%	5.15%	3.89%
3	1.74%	3.37%	2.78%	3.44%	2.46%

资料来源：Wind，兴业证券经济与金融研究院整理

那么为什么使用卡尔曼滤波能够有效提升基金仓位测算的准确性呢？从原理上看，卡尔曼滤波与传统回归方法最大的不同在于它是一种时间序列方法，使用了贝叶斯的思想。我们将基金行业仓位当成一种隐含状态，把基金收益率当成可观测变量，卡尔曼滤波在迭代时会比较基于模型得到的基金收益率预测值与当期实际基金收益率的差异，差异越大说明之前估计的基金仓位越不准确，因此当期需要对仓位做更大程度的调整，反之则无需做大幅调整。调整的最优幅度就由卡尔曼滤波给出。实际上从原理看，卡尔曼滤波法不仅可以用于基金行业仓位的估计，它对于基金久期、基金经理风格识别等领域都有潜在应用价值。

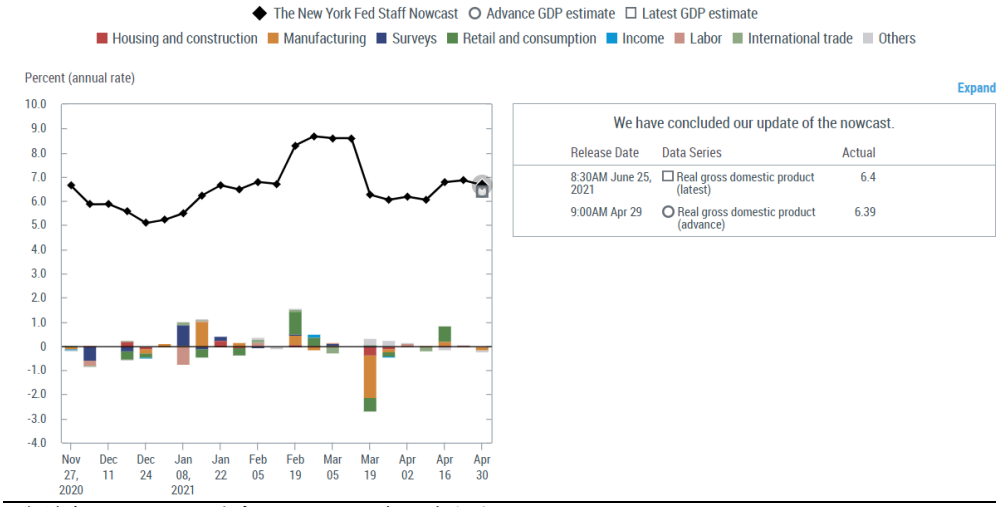
3.卡尔曼滤波与宏观数据实时预测

卡尔曼滤波在宏观经济数据的建模和预测中也能够发挥重要作用。对于政府机构来说，当前经济状态直接影响到宏观政策的制定。例如，央行需要判断当前经济是否过热，从而决定是否需要通过加息等措施来给经济降温。对于投资者来说，经济状态的变化也直接影响投资策略。

而宏观数据的特征使其处理较为困难，难以用简单模型如线性回归来对其建模和预测。举例来说，首先宏观数据公布时间往往不同步，且可能有缺失值；其次宏观数据频率不一致，有季频、月频等等；并且宏观数据截面维度高，但时间序列长度较短；那么如何综合利用这些特征各异的大量宏观数据进行建模，识别新数据发布带来的边际冲击，并获得经济状态的及时更新呢？

基于动态因子（DFM）的实时预测模型是识别当前经济状态最先进的方法之一。它的思想就是利用所有可得的信息来获得当前经济状态的估计，是一种状态空间模型。举例来说，GDP 增速是反映经济状态的最直接指标，但由于它只在每季度发布一次，且发布时间滞后，因此我们总是无法得到当前季度的 GDP 增速。这样一来，我们就必须使用其他更高频指标如工业增加值增速等来推测当前季度的 GDP 增速。比如纽约联储维护了一个美国 GDP 环比的实时预测模型，会在有新宏观数据发布后利用卡尔曼滤波更新预测结果。下方给出了一个对 2021Q1 美国 GDP 环比预测过程的示例。可以看到随着数据的发布，预测值在不断调整，最终预测值为 6.66%，与 6.4%的实际发布值非常接近。

图表 8、纽约联储 Nowcast 模型 2021Q1 预测示例



资料来源：纽约联储官网，兴业证券经济与金融研究院

利用状态空间模型，我们可以将大量宏观数据的变动解释为少量隐含因子的变动，从而解决数据截面维度高的问题。在估计得到模型参数后，我们就可以通过卡尔曼滤波来估计隐含因子，然后通过测量方程（1）得到所有缺失宏观数据的

预测值了。标准的实时预测过程是在每个时点获取所有可得数据，然后据此拟合DFM模型、更新预测结果。这里我们简要说明卡尔曼滤波是如何实现缺失值的处理与预测的。假设我们使用等式（1）（2）代表的状态空间模型来对宏观数据 z_k 建模，并用若干隐含因子 x_k 来解释宏观数据的变动:

$$\begin{cases} z_k = Hx_k + \eta_k, \quad \eta_k \sim N(0,R) \\ x_k = Ax_{k-1} + e_k, \quad e_k \sim N(0,Q) \end{cases}$$

其中 η_k 为服从正态分布、协方差矩阵为 R 的白噪声过程;e_k 为服从正态分布、协方差矩阵为 Q 的白噪声过程。

在上述模型中，假设数据 z_k 没有缺失值，是前后完整对齐的数据，那我们直接通过卡尔曼滤波可以估计出隐含因子 x_k，并利用转移方程向后预测。但实际中，宏观数据 z_k 往往是不对齐的，比如有的宏观数据历史数据长，有的则较短，导致数据矩阵的前段不对齐。另外，宏观数据 z_k 的发布时间可能不同，导致在某一时刻的数据矩阵后段也不对齐。再考虑到比如中国一些数据在1月不发布等问题，数据矩阵的中间也会有空缺。也就是说我们面临的是一个“千疮百孔”的宏观数据矩阵。而卡尔曼滤波可以轻松的处理这些缺失值，实现隐含因子的估计，并自动填充缺失值。

假设数据 z_k 有缺失值，我们用矩阵 W_k 来表示缺失值的位置，即 W_k 是在 z_k 数据缺失行取值为0的单位阵。我们将模型形式调整如下:

$$\begin{cases} z_k^* = H^* x_k + \eta_k^*, \quad \eta_k^* \sim N(0,R^*) \\ x_k = Ax_{k-1} + e_k, \quad e_k \sim N(0,Q) \end{cases}$$

其中 $z_k^* = W_k z_k$，$H^* = W_k H$，$\eta_k^* = W_k \eta_k$，$R^* = W_k R W_k^T$。在卡尔曼滤波的迭代过程中，我们只需要在遇到缺失值时使用新的模型形式就可以毫不费力的估计出隐含因子 x_k 了。然后，基于原始测量方程，我们只需要用测量矩阵 H 乘以 x_k 就得到模型拟合的完整 $\hat{z}_k$ 向量，从而实现缺失值的填充。实际上，在实际预测宏观数据的过程中，我们就是将需要预测的目标数据点看作缺失值，统一利用卡尔曼滤波填充和预测。

使用动态因子模型的另一个特性是可以将新数据发布对预测值的影响程度分解。假设基于某时刻可得的数据集 Ω_1，模型对于目标宏观数据的预测值为 a_1，而一段时间后若干新数据发布，基于新数据集 Ω_2 得到的预测值为 a_2。则每个新数据对目标宏观数据预测值的影响程度都满足如下等式:

信息冲击=权重*预期差=权重*（公布值-预测值）

下方图表给出了纽约联储在 2021 年 4 月 16 日对新发布数据信息冲击的分解示例。可以看到，基于 4 月 9 日数据集的 GDP 环比预测值是 6.05%，而由于新数据发布预测值修订为 6.78%，增加了 0.73%。其中 Retail Sales and Food Services 数据的发布值为 9.82，信息冲击（Impact）为 0.65%，即 0.73%的增加值中有 0.65%是超预期的零售数据贡献的。

图表 9、纽约联储 Nowcast 模型的数据发布信息冲击分解示例

Data Flow (Apr 16, 2021)

Model Update	Release Date	Data Series	Actual	Impact	Nowcast GDP Growth
Apr 16					**6.78**
	8:30AM Apr 16	Building permits	46.00	-0.02	
	8:30AM Apr 16	Housing starts	19.35	0.06	
	9:10AM Apr 15	Capacity utilization	1.04	0.08	
	9:10AM Apr 15	Industrial production index	1.44	0.07	
	8:30AM Apr 15	Philadelphia Fed Mfg. Business Outlook: Current activity	50.20	-0.01	
	8:30AM Apr 15	Empire State Mfg. Survey: General business conditions	26.30	0.00	
	8:30AM Apr 15	Retail sales and food services	9.82	0.65	
	8:30AM Apr 14	Export price index	2.14	0.02	
	8:30AM Apr 14	Import price index	1.25	-0.01	
	8:30AM Apr 13	CPI-U: All items less food and energy	0.34	0.01	
	8:30AM Apr 13	CPI-U: All items	0.62	0.01	
		Data revisions		-0.13	
Apr 09					**6.05**

资料来源：纽约联储官网，兴业证券经济与金融研究院

值得注意的是，动态因子模型是一种纯粹基于数据的计量模型，它的准确与否取决于数据的质量和更新的速度。另外，相比模型给出的预测绝对值，新信息带来的预测值边际变化可能对于指导投资有更重要的意义。通过此模型，我们将大量的宏观数据融合起来，能够实现系统化的经济状态和数据发布监测。

4、总结

卡尔曼滤波(Kalman Filter)是一种利用线性系统状态方程，通过系统输入输出的可观测数据，对系统隐含状态进行最优估计的算法，可以从受误差影响的传感器测量中估算出最佳的实际系统状态，常用于制导与导航控制系统、计算机视觉系统和信号处理等领域。本文创新性的发现卡尔曼滤波在投资研究中也能够产生重要作用。

我们把基金的行业仓位当做隐含状态，把基金收益率当做可观测状态，然后用状态空间模型对基金收益率建模，那么通过卡尔曼滤波可以高效的识别出基金持仓的变动。与传统基于线性回归的方法相比，卡尔曼滤波是一种基于贝叶斯思

想的时序方法，行业平均仓位的估计误差能够减少 30%以上，重仓行业仓位估计误差减少约 50%，对于指导投资构建交易策略具有较高价值。

在宏观数据建模和预测领域，通过使用动态因子模型建模，卡尔曼滤波可以实现宏观数据共同推动因子的估计，并能够实时测算新发布数据对 GDP 等重要数据预测值的信息冲击程度。

除了本文提到的基金行业仓位测算和宏观数据建模，卡尔曼滤波在基金久期测算、基金经理风格识别等领域也有潜在的应用价值，我们后续将继续探索下去。

5、参考文献

[1] Simon, D. (2010). Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. IET Control Theory & Applications, 4(8), 1303-1318.
[2] Gupta, N., & Hauser, R. (2007). Kalman filtering with equality and inequality state constraints. arXiv preprint arXiv:0709.2791.
[3] Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. Annual Review of Economics, 10, 615-643.
[4] Bańbura, M., & Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. Journal of Applied Econometrics, 29(1), 133-160.

风险提示：模型结论基于历史数据，在市场环境转变时模型存在失效的风险。

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

投资评级说明

投资建议的评级标准	类别	评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级(另有说明的除外)。评级标准为报告发布日后的12个月内公司股价(或行业指数)相对同期相关证券市场代表性指数的涨跌幅。其中：A股市场以上证综指或深圳成指为基准，香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	相对同期相关证券市场代表性指数涨幅大于15%
		审慎增持	相对同期相关证券市场代表性指数涨幅在5%～15%之间
		中性	相对同期相关证券市场代表性指数涨幅在-5%～5%之间
		减持	相对同期相关证券市场代表性指数涨幅小于-5%
		无评级	由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级
	行业评级	推荐	相对表现优于同期相关证券市场代表性指数
		中性	相对表现与同期相关证券市场代表性指数持平
		回避	相对表现弱于同期相关证券市场代表性指数

信息披露

本公司在知晓的范围内履行信息披露义务。客户可登录 www.xyzq.com.cn 内幕交易防控栏内查询静默期安排和关联公司持股情况。

使用本研究报告的风险提示及法律声明

兴业证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告仅供兴业证券股份有限公司（以下简称“本公司”）的客户使用，本公司不会因接收人收到本报告而视其为客户。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本报告所载资料的来源被认为是可靠的，但本公司不保证其准确性或完整性，也不保证所包含的信息和建议不会发生任何变更。本公司并不对使用本报告所包含的材料产生的任何直接或间接损失或与此相关的其他任何损失承担任何责任。

本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌，过往表现不应作为日后的表现依据；在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告；本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现。过往的业绩表现亦不应作为日后回报的预示。我们不承诺也不保证，任何所预示的回报会得以实现。分析中所做的回报预测可能是基于相应的假设。任何假设的变化可能会显著地影响所预测的回报。

本公司的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

本报告并非针对或意图发送予或为任何就发送、发布、可得到或使用此报告而使兴业证券股份有限公司及其关联子公司等违反当地的法律或法规或可致使兴业证券股份有限公司受制于相关法律或法规的任何地区、国家或其他管辖区域的公民或居民，包括但不限于美国及美国公民（1934 年美国《证券交易所》第 15a-6 条例定义为本「主要美国机构投资者」除外）。

特别声明

在法律许可的情况下，兴业证券股份有限公司可能会持有本报告中提及公司所发行的证券头寸并进行交易，也可能为这些公司提供或争取提供投资银行业务服务。因此，投资者应当考虑到兴业证券股份有限公司及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。

兴业证券研究

上海	北京	深圳
地址：上海浦东新区长柳路36号兴业证券大厦15层	地址：北京西城区锦什坊街35号北楼601-605	地址：深圳市福田区皇岗路5001号深业上城T2座52楼
邮编：200135	邮编：100033	邮编：518035
邮箱：research@xyzq.com.cn	邮箱：research@xyzq.com.cn	邮箱：research@xyzq.com.cn